Network statistics on early English Syntax: Structural criteria

Bernat Corominas-Murtra[1]

[1] ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain

Abstract

This paper includes a reflection on the role of networks in the study of English language acquisition, as well as a collection of practical criteria to annotate free-speech corpora from children utterances. At the theoretical level, the main claim of this paper is that syntactic networks should be interpreted as the outcome of the use of the syntactic machinery. Thus, the intrinsic features of such machinery are not accessible directly from (known) network properties. Rather, what one can see are the global patterns of its use and, thus, a global view of the power and organization of the underlying grammar. Taking a look into more practical issues, the paper examines how to build a net from the projection of syntactic relations. Recall that, as opposed to adult grammars, early-child language has not a well-defined concept of structure. To overcome such difficulty, we develop a set of systematic criteria assuming constituency hierarchy and a grammar based on lexico-thematic relations. At the end, what we obtain is a well defined corpora annotation that enables us i) to perform statistics on the size of structures and ii) to build a network from syntactic relations over which we can perform the standard measures of complexity. We also provide a detailed example.[1]. Keywords: Syntax, complex networks, learning, Computation

[1] *This paper is the experimental design of a more extensive work* **The ontogeny of syntax networks through Language Acquisition**, *Corominas-Murtra, B., Valverde, S. and Solé, R. V.*

Contents

1 Introduction 3
1.1 Different abstractions, different questions: Syntax and Statistical Physics . . . 3
1.2 Aims . . . 4

2 Syntactic networks 4
2.1 From syntax to networks: what we win and what is lost . . . 4
2.2 Syntactic Networks . . . 5
2.2.1 From syntactic relations to links . . . 5

3 Data 8

4 Building the Networks of Syntactic Acquisition: Criteria 9
4.1 Non accepted productions . . . 10
4.2 Accepted Productions . . . 10
4.2.1 Phrases and missing arguments . . . 11
4.2.2 *To be* verb . . . 11
4.2.3 Infra-specification and semantic extension of lexical items 13
4.2.4 First functional particles . . . 14
4.2.5 Duplication of functional words . . . 16
4.2.6 Non-structural lexical items . . . 17
4.2.7 Negation Structures . . . 18

5 Corpora Annotation 19

6 The average size of structures, $\langle \mathcal{S} \rangle$ 19

7 Building the Network 20
7.1 Measures . . . 20

8 Example 22
8.1 The source . . . 22
8.2 Selected Productions and Analysis . . . 23
8.2.1 XML Format to be read by DGAanotator . . . 24
8.2.2 Selection of *valid* strings, annotation and computation of $\mathcal{S}$ 25

9 Acknowledgments 26

1 Introduction

In this pages there is an attempt to design and describe a naturalistic experiment on syntax acquisition. Specifically, we want to build a *Syntactic network* in order to study syntax with modern methods of complex network theory. The process is nor standard neither straightforward and deserves to be well described.
There are interesting descriptive frameworks based on networks to study syntax. One of them is the so-called *Dependency grammar*[1]. There are, also, theoretical approaches using graphs. A remarkable member is the *word grammar*[2]. The approach assumed here is closer to the Word-Grammar, despite we develop our own criteria, as well as we consider the graph representation as a linear projection of the constituency hierarchy.
The paper is organized as follows: We firstly discuss the scope and validity of the conceptualization of syntactic relations within a network. The core of the work is devoted to the discussion of the (descriptive) structural criteria to tackle the problem of annotation in early grammars. Finally, a brief compendium of network measures is shown, as well as an illustrating example. All analysis are performed over the PETER corpora of CHILDES database [3] using the DGA-Annotator [4].

1.1 Different abstractions, different questions: Syntax and Statistical Physics

Every abstraction of a natural object implies a particular conception of it in order to answer a specific question. Assuming that every abstraction implies a simplification, we have to explore, then, how different approaches can be complementary or whether some of these approaches are more fruitful than others -i.e., what are the core questions leading to the understanding of such phenomena. Focusing on language, research on syntax seeks to find the minimal set of rules that could generate all -and only- the potentially infinite set of sentences of a given language. Thus, the question addressed by syntax is the problem of decidability or computability of the set of possible sentences of a given language. When dealing with language as a complex network, we have to note that statistical physics works from different perspectives: What are the global features of the dynamics of our system? How the combinatorial space is filled? What is -if any- the role of constraints?
Thus, we don't address questions concerning the structure of the inhabitants -sentences- of our system, but its global dynamics and organization. Note that the questions are different than in the case of syntax: thus, the abstraction we are working in is also different. Note, also, that we are not negating nor denying the particular features of sentence construction. Simply, we work at other level of abstraction. We are confident that information from this different level of approach should be enlightening to questions addressed on grammar itself.
If one wants to apply statistics on some syntactic phenomena, a word of caution is needed because there is a gap between the syntactic procedure and the statistical physics procedure: The former is focused on explaining almost *every*

subtlety of sentence construction, while the latter works on averages over the largest possible set of data. Thus, a compromise has to be assumed because it is not possible to deal with every syntactic phenomena but, also, the statistics has to be built on certain criteria.

1.2 Aims

Thus, the aim of this document is to present a set of descriptive criteria to identify *structure* in early child grammars. This is not a theoretical reflection about the concept *structure* or its evolution during the process of language acquisition. With these criteria, we want to build up the so-called syntactic networks from early child grammars. Indeed, even though many features of the language acquisition process have been identified and well studied, there is a lack of a clear concept of what it is structured or not in early child grammars, namely, there is not a concept such as grammaticality [5] or *convergence* [6], defined in adult grammars. If we take the adult-grammar concept of grammaticality, we will surely reject almost all of children's productions. But it will not be true that many of these rejected utterances are unstructured at all.
In order to overcome these limitations, we developed a set of descriptive criteria to extract the syntactic network of different sets of the child's utterances belonging to successive time stages of the language acquisition process. As we discussed above, we present these criteria employed in the construction of the associated networks[2].

2 Syntactic networks

2.1 From syntax to networks: what we win and what is lost

Formally speaking, a given language, $\mathcal{L}$ is composed of an arbitrary large, but finite set of lexical items $\mathcal{W}$ - or alphabet, in technical words- and of a restricted set of rules and axioms, Γ. These rules describe how the elements from $\mathcal{W}$ can be combined in order to *1.* obtain sentences of $\mathcal{L}$ or to *2.* decide if a given sequence of elements from $\mathcal{W}$ is a sentence of $\mathcal{L}$ or not [7]. Syntax properties are, thus, indicators of grammatical complexity[3]. If one intends to develop a syntax theory to decide whether a given sequence of words -namely, Russian words- is a sentence of *Russian* language, one needs to develop rules involving hierarchical and long range relations. Moreover, the set of rules must be *generative*, in the sense that they should involve some recursive condition to grasp the potential infinity of sentences generated by Russian grammar [9].

[2]Note that many properties of the networks make sense asymptotically, i.e., many utterances need to be analyzed such that the results acquire statistically significance.

[3]In fact, if we would be able to design the minimal program to describe our system, its size (in bits) would be an index of complexity. See [8]

Thus, we can say, without any loose of generalization, that syntax works at the *local level* of language[4], i.e.it operates at the sentence level, no matter how long the sentence is. Now, we wonder about the global profiles of syntactic relations. Note that the question we want to address is not finding the specific rules needed to generate the possible sentences of $\mathcal{L}$, but we want to take a look at the system as a whole. This could seem bizarre when considered from the point of view of mainstream theories of syntax, but it is a common procedure in statistical physics. Global profiles can provide information about general dynamics and constraints acting over the whole system as a complex entity. The unexpected profile given by Zipf's law is an example of global behavior of language dynamics [10].
A note must be added concerning the naturalistic character of this kind of experiments. Syntax has been related with competence abilities. But statistical and naturalistic works are carried out over performance data. Thus, we are inferring the global patterns of performance by assuming some competence abilities.

2.2 Syntactic Networks

Networks revealed as an interesting abstraction to explore the global behavior and dynamics of complex real systems made from units and the associated relations between such units. Let's explore such abstraction for syntax relations. A network $\mathcal{G}(V, E)$ is defined by the nodes V and the links E relating the nodes V [11]. These links can be directed or undirected; we will use the directed ones, if the contrary is not indicated. To build a syntactic network, the mapping of $\mathcal{L}$ onto a graph will be straightforward for the set $V \rightarrow \mathcal{W}$ i.e., the set of lexical items of $\mathcal{L}$ will be the set of nodes of $\mathcal{G}$. The mapping from Γ to E is not so obvious and needs further considerations.

2.2.1 From syntactic relations to links

As we discussed above, the syntactic rules needed to generate any natural language revealed considerable degree of complexity. Thus, it is clear that the statistical treatment employed here is an approximation. Modern syntax is based on recursive operations of *merge* and *move* [6]. Such operations lead the syntactic derivations to display hierarchies and long range relations. These are features that cannot be captured explicitly by a descriptive framework based on linear relations among lexical items -a network approach. But we are approaching the language structure from the point of view of statistical physics: we want to capture the global patterns of the system, thus we cannot specify *all* the local properties. This is contrary to the procedure employed in the Ising models of ferromagnetism, despite the success of this approach is universally acknowledged. Thus we have to decide what is the most essential structure in a

[4]We are not considering the usual *locality* of syntactic relations as understood in many works of syntax, we use the term *local* to specify that syntax operates at the level of individual elements of a given language $\mathcal{L}$

syntactic derivation. We assume that the most fundamental thing one can say from the syntactic point of view about a sentence is its *constituent structure*. Constituent structure can be captured by linear relations. In the following, we define an exact mapping from a hierarchical binary tree to a graph[5], an entity made of binary relations (see figure (1)):

1. Find the basic syntactic structure of constituents without labels nor internal operations, with clear distinctions of complements and the head in every phrase. Detect the verbs in finite forms.

2. Trace an arc from the complement to the head of the phrase. If the complement of a given phrase is also a phrase, trace and arc from the head of the internal phrase to the head of the external phrase. We want to recover the merging order.

3. The head will be the semantically most relevant item.

4. The verbs in finite forms are the head of the sentence.

With the above criteria, we make an attempt to manage data with the less aggressive criteria. Moreover, these assumptions don't constrain us to one or other linguistic school and grasp reasonably with the observed syntactic development of children.
With this method, we do not restrict our set of sentences to the one generated by finite combinatorics. We allow our sentence to be arbitrary long. Thus, our model is only finite because real data is finite in nature, but it doesn't negate the theoretical possibility of infinite generativity, a property expected for any approach of syntax [12]. Moreover, the relevance of the statistical physics properties generally is found in systems asymptotically large.

[5]In the approach of Word-Grammar, the projection of hierarchical structures into linear dependencies is just the inverse of what we adopted here. But it is, essentially, the same procedure. For more information, see [2]

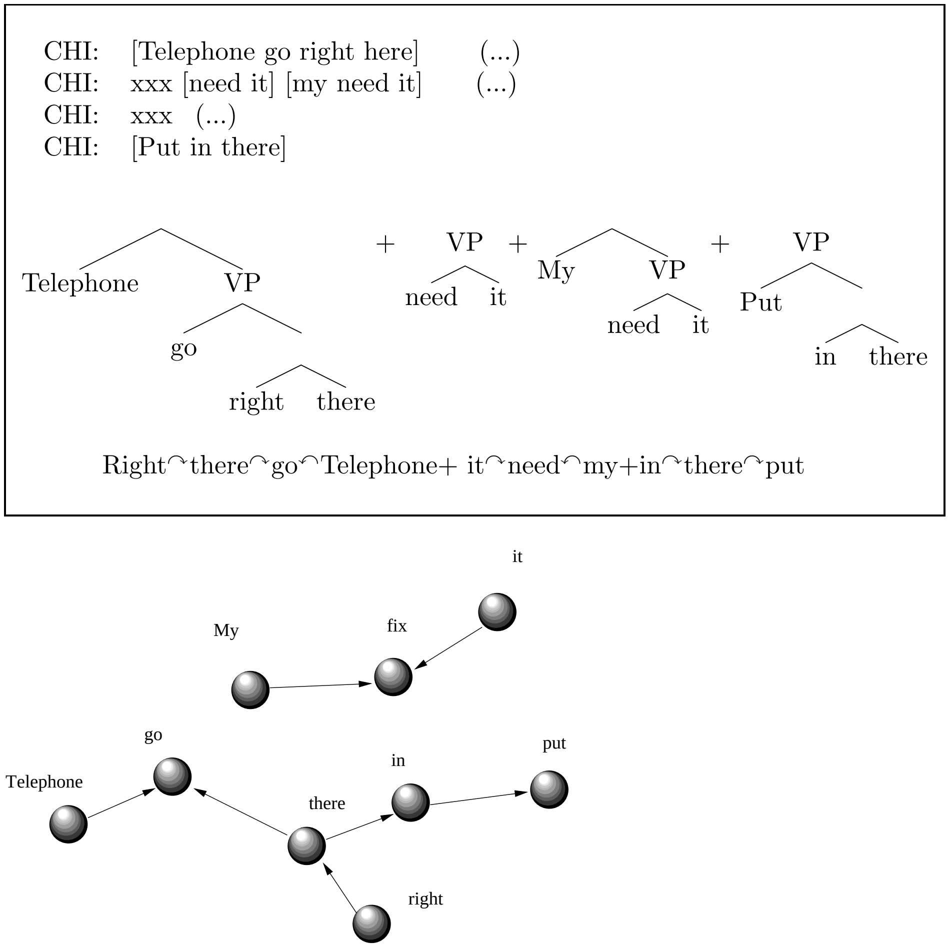

Figure 1: *Building syntactic nets from children free speech corpora.* **A**) *We have the transcript of a conversation and we select only child's productions. We identify the structured strings . The notion of structure and the used criteria is widely developed in further considerations.* **B**) *Basic analysis of constituent structure, identifying the verb in finite form (if any) in different phrases.* **C**)*Projection of the constituent structures into lexical dependencies (note that the operation is reversible: We can rebuild the tree from the dependency relations.).* **D**)*Following the dependency relations found by projecting the naked syntactic structure we build, finally, the graph.*

Syntactic networks can be built by other procedures. Dependency syntax [1] has been used in other works [13]. In such an approximation, syntactic networks have been built up by assuming syntactic relations as dependency relations among lexical items. Dependency grammar generates a graph to describe the sentence structure and it is the reason why it is interesting to build networks.

The mechanism to build large nets is straightforward[6].

3 Data

Studies on language acquisition can be divided into two main types: *experimental* and *naturalistic*. The experimental ones are focused on child's response to well-established situations in order to obtain data of some specific trait. Naturalistic studies, at the other hand, are based on child's free speech corpora. These corpora can be extracted, for example, from a recorded session where the child speaks with adults spontaneously. [15]. Our study is clearly naturalistic, and this label takes here its whole meaning, because the procedures to build up *biological* networks, for example, are conceptually, the same.
Data has been extracted from the well-known *CHILDES Database*[7] [3, 16, 17]. The chosen corpus is the *Peter* Corpus, from Bloom 1970. We choose this data for many reasons: 1) Time intervals are regular (about 2 or 3 weeks). 2) Extension of the corpora can be considered large enough to seize global properties, taking into account the intrinsic small size of the system. There is a little exception in corpus 2, which is, by far, the smallest one. Fortunately, this corpus does not seem to belong to a key stage in grammar evolution. 3) The acquisition stages of Peter seem to be the standard ones observed in language acquisition. Thus, it is reasonable to think that our results will not be biased to strange deviations of the particular case study.
Working data includes the 11th first corpora of Peter's 20 corpora. The age period goes from 1 year and 9 months to 2 year and 4 months. As we said above, the aim of the study is to observe whether and to what extent syntactic networks can provide information on the process of language acquisition. The so-called *syntactic spurt* ([15]), which appears later than the *lexical spurt*, is clearly observable in the chosen corpora. Thus, we manage data that begins when the lexical spurt has already taken place and ends when syntactic structures of child's productions are complex enough to be compared with the adult ones. This does not exclude the possibility of more abrupt changes in more advanced acquisition stages, but we stop our analysis here.

Material contains several conversations between adults and the child (These adults are, mainly, researchers and Peter's parents). We selected the child's productions and we studied them considering the discursive context where such

[6]some authors assume the network abstraction for syntax as ontological, i.e., not as an approximation to a complex system of rules involving recursive structures and non-terminal nodes(see [13], [14]). This is not the view adopted here: The network in our approach only is an attempt to grasp some evolutionary features of the system, properties that can be captured by taking a global view to the system, something that is difficult to achieve when looking at the local structure of syntactic relations. Here, networks do not substitute the decision/computation rules because some key features of the syntax itself, such as constituent hierarchy or movement, cannot be treated properly by the graph theoretic abstraction

[7]http://talkbank.org

Corpus	**Age**	**Corpus**	**Age**
1	1;9.7	7	2;1.0
2	1;9.21	8	2;1.21
3	1;10.15	9	2;2.14
4	1;11.7	10	2;3.0
5	1;11.21	11	2;3.21
6	2;0.7		

Table 1: Age of Peter in successive corpora (**years;months.days**). Data from childes database [3, 16, 17]

utterances have been produced. This enables us to *clean* the data. What it means is that we will discard 1) imitations from adults 2) non-structured utterances. A complete explanation of the criteria to accept productions and by this implying that they contribute to the syntactic graph is reported in the next section *Criteria*.
A final note concerning the data: it seems clear that the morphological nature of English, with poor inflectional features makes the identification of functional items easier than in a language with richer inflectional features. The global impact of the morphological nature of a given language on network topology cannot be denied [18], but the global reorganization process observed in child syntactic networks seems to go beyond these singularities[8].

4 Building the Networks of Syntactic Acquisition: Criteria

We selected the productions that allow us to identify some syntactic structure. Obviously, the word *criteria* is due to the evidence that despite the fact that most of early child-productions are not grammatical in the sense of full convergence or complete feature checking, it is not true that they have no structure. Thus, the work of the linguist consists in identifying the clues of syntactic structure in child's productions. Selection is not easy at all, as there does not exist an explicit definition of syntactic structure in early grammars. We considered that there exists structure if there exists, at least, some lexico-thematic relation between the elements in a production. This is the basis of syntactic structure of early English grammars [15]. More complex relations, involving functional words, appear later and syntactic structure can be more easily identified. This is coherent with the observed nature of early grammars.

[8]obviously, words as fundamental units is an intuitive but rather arbitrary choice. Thus, the same study could be extended by considering morphemes as the fundamental unit. This is, maybe a more reasonable choice. In this way, it could be possible to detect more similarities when comparing the acquisition processes of different languages.

4.1 Non accepted productions

First of all, we discarded some transcribed strings if: 1)they are simply an onomatopoeia with no structural role (in some cases *choo choo* could replace *train*). 2)they are non transcribed items -because we supposed it was not possible to understand what the child said. We choose not to consider any of these unidentified lexical items (transcribed in the corpora as *xxx* or *yyy*) in order to ensure the transparency of data managing.

These non-accepted elements are: *a* (in some specific contexts), *ah, an* (in some specific contexts), *awoh, ay, hey, hmm, huh, ka, ma, mm, mmhm, oh, oop, oops, ow, s* (in some specific contexts) *sh, ssh. ta* (in some specific contexts) *uh, uhhuh, uhoh, um, whoops, woo, yum.* **Onomatopoeia**: *choo, Moo, Woof, Bee Bee*

The case concerning *a*, the schwa, will receive a particular attention below.
Some onomatopoeia appear together with its corresponding lexical item. To analyze them, we assume onomatopoeia to be nonexistent. Take, for example:

Peter 9*I want ta write the choo choo train* → *I want ta write the train*

Considerations related to other non-trivial interpretations, such as the role of *ta*, are extensively developed in the following lines.
In addition, and following the enumeration of non-accepted productions, we find the general case where no structure is identified in a production. In this situation, we consider the utterance as a string of isolated lexical items. Consequently, no links but only nodes corresponding to the lexical items are added to the graph.
More attention has to be paid to imitations. The reason to consider imitations as unacceptable productions is that we have no confidence that such string is identified as a structured one or, simply, as a single lexical element. Imitations are identified by analyzing the discursive context. Some utterances of surprising complexity for its corresponding stage are produced after an *untranscribed adult conversation*: we cautiously removed from the graph such contributions. In Peter 5 corpus, *I can't see it* is produced after an adult conversation and it is, by far, the most complex production of this corpus. It strongly suggests that this is an imitation from something said in such untranscribed adult conversation.

4.2 Accepted Productions

As we stated above, structured productions and lexical items are taken into account. Now we state another assumption: If in the whole utterance we cannot

find global structure but there are some structured strings, then we take these structured strings separately, (see figure 1).

4.2.1 Phrases and missing arguments

In the pre-functional stage (identified in our data until corpus Peter 6) there appear a lot of utterances where only thematic relations seem to be considered by the syntactic system of the child. Thematic relations are fundamental at the syntactic level, and their appearance indicates presence of sub-categorization mechanisms in child grammar. No traces of more complex structure -like agreement- is found in this early stage of acquisition. We consider as syntactic relations the thematic relations between verb and arguments. Moreover, subject elision is usual, due mainly to the facts that 1)utterances are in imperative mode or 2)there is no fixation yet of parametric variation associated to the explicit presence of subject in English. Productions of this kind are:

Peter 5 *Open box*, instead of *Open the box*, (the determiner is missing.)

Peter 5 *wheel walk* instead of *The wheel walks*, (3-singular English agreement is missing)

Peter 6 *two truck* instead of *two trucks*, (no plural agreement)

This leads to the logical conclusion that productions like **open the* will not be accepted. The reason is clear: if we assume thematic relations as the basic building blocks of child syntax, the non-presence of the semantically required argument but its determiner is not enough to define any relation.
Relations between verbal head and functional words are specifically considered in phrasal verbs. Its isolated production is considered a structured utterance. Several reasons support our choice: 1) Their intrinsic complex nature, 2) We cannot conclude that there are lexicalized imitations because, in adult speech, phrasal verbs usually are *broken* by a noun or determiner phrase:

Turn [the wheel]$_{SD}$ out.

4.2.2 *To be* verb

Semantically vacuous predications (those which involve the *to be* verb) are often produced without realization the verb. We argued that missing arguments or lack of agreement in a production could be not the only reasons to conclude that there is not any structure in child utterances. This was justified because strong semantically items were present in discussed productions. The case of copulative constructions will be treated close to the ones involving missing functional words. In this case, no presence of the verb does not motivate the consideration

of non-structured production. An interesting production is:

Peter 5 *Wheels mine* Instead of **The** *wheels* **are** *mine*

In this case we have a predication *mine* from something *Wheels*. Formally, *are* is a semantic link between predication and the element from which something is predicated [**?**]. So the missing of the *to be* verb could be considered analogous to the missing of a functional particle. The same situation arises from:

Peter 7 *That my pen*

Usually, when inflectional morphology appears, some infinite forms are present without the finite form of the *to be* verb. This is the case of some present continuous utterances such as:

Peter 8 *I writting*

This case should be treated as the above case: There is some predication with semantic structure. Just the opposite is also found: presence of the *to be* verb with an infinitive or finite form:

Peter 8 *I'm write too*

In this case, we could assume that the child is acquiring inflectional morphology and that this utterance is a present continuous one without inflection. In the other hand, we could consider that *'m* has not a role in the sentence and thus, this can be treated as a single finite sentence *I write too.*
Analogously,

Peter 7 *I'm do it*

or

Peter 6 *cars goes away*

Are treated as single finite sentences: *I do it* and *The cars go away*
Some lexicalized phrases in adult language, such as *back seat* or *thank you*, have been considered as a complex structures. The reason is to be coherent: If we assume that *fix it* is clearly an imperative structured sentence, at this stages of acquisition there is no reason to think that *back seat* or *thank you* have to be considered differently. Moreover, this interpretation is also coherent with the one developed for phrasal verbs.
A special case of imitations involving the *to be* verb will be accepted. These imitations involve some *adaptation* of adult syntax to the syntax in which the child is competent. An example should be:

(Peter 6)

Adult: *Is that a truck?*

Child(1): *That's a truck?*

Child(2): *That a truck?*

In this example, adult production involve an interrogative sentence with subject inversion. The first imitation (*Child*(1)) retains all the lexical items but the sentence is translated as interrogative without subject inversion. In the second successive imitation (*Child*(2)) the verb is missing. But the elements to define a predication are still at work -with a *schwa* as a determiner, suggesting that the child is entering into the functional stage.

4.2.3 Infra-specification and semantic extension of lexical items

During the acquisition process, extension of meaning is subject to variations. To know which is the intrinsic nature of these changes is not our aim, but we have to manage such situations. Thus, we find utterances where the child uses in the *wrong* way some lexical item that could be related semantically with the *right* lexical item. As an example:

Peter 5 *More screwdriver*

Which could, checking the context, be properly replaced by constituents or lexical items with related meanings:

Another screwdriver

or

screw it again

or

screw it more (or *harder...*)

In the first case, we could consider that the child made some semantic extension of the word *more* and it has enough traces to define a syntactic relation. But context can lead us to a second or third interpretation. Generally, if there is a great ambiguity we reject such utterances as structured ones. In this case, we should not consider any syntactic structure. Thus, we don't define any relation in productions such as:

Peter 5 *Screwdriver help*

Peter 5 *More [fix it]*$_{SV}$ (We don't define any relation between *More* and the SD *fix it*)

The semantics of the productions are intuitive, but is hard to justify clearly some kind of syntactic dependency.
Strings displaying mistakes in the use of personal pronouns and possessives have considered as structured. Generally, we could associate such mistakes to the absence or weakness of case system. But many productions, as we reported above, have structure without any trace of case assignation. Examples of this kind of utterances are:

Peter 5 *My fix it* instead of *I fix it*

Peter 8 *Me write* instead of *I write*

This assumption is reinforced by realizing that, in some cases, a production with wrong pronoun is repeated correctly without any conversational pause:

Peter 8 *Me found it (...) I find it*

This situation cannot be confused with the missing of the *to be* verb such in the case of *wheels mine*. This case has to be considered as above mentioned when dealing with missing *to be* verb structures.

4.2.4 First functional particles

In early corpora (1-4) child productions display very poor structures. This is the so-called *pre-functional* stage, where no functional words appear in structured productions. Beyond this point, some lexical elements -we are mainly talking about the *a*, the *schwa*- seem to act as a *proto*functional particles. Whether this schwa has a phonological or functional-syntactic character is an open question [19, 20].
Some authors related the presence of these items as one step to combinatorial speech [20], but they realized that, in early stages, the role of these items is more related to phonological processes of language acquisition, without any functional or structural role, at the syntactic level. Other authors such as Veneziano & Sinclair "*linked these phenomena more specifically to the child's development of grammatical morphemes considering them as a sort of an intermediate form on the way to grammatical morphemes.*" [19]pp 463. Roughly speaking, we can say that the core of this reasoning is rooted in the idea that the role of such items is dynamic, going at very first stages as *filler syllabes* without any syntactic role and acquiring grammatical features during the process to end as functional particles, with specific syntactic role.
The lack of consensus around a topic that seems to be crucial in syntactic acquisition theorizations forces us to be really cautious in interpreting such items. Furthermore, functional words such *a* are strongly candidates to be the hubs in a fully developed syntactic network. Hub are the most connected nodes on a network, being, thus core pieces in network organization. Every candidate

belonging to the set of functional particles is specially analyzed in order to discard simple phonological phenomena. Thus, for every occurrence of such items there will be an individual decision, taking in account the context and with the framework defined by Veneziano & Sinclair [9].
Specifically, we considered that sometimes the *schwa* plays a functional role. It is reasonable to assume, thus, that sometimes the *schwa* is substituting a specific functional particle. In this cases we assume that the *schwa* acts within the syntactic structure as the substituted particle. Several examples can illustrate such reasoning:

Peter 6 *Light* **a** *hall*

Peter 6 *light in* **a** *hall*

Peter 6 *look* **a** *people*

In this case, it seems that *a* substitutes *the*. **a** should be treated as a determiner. This is a very difficult choice, because purely phonological interpretation could be enough to justify the presence, specially in the third case.
Sometimes choice is really ambiguous. Take for example:

Peter 6 *There* **a** *new one*

Such a case **a** could be easily interpreted as a pure phonological phenomena. But if we consider the vacuous semantic nature of the *to be* verb, we could understand these occurrences as protofunctionals. We removed these most ambiguous cases. We also rejected as unstructured utterances productions involving confuse sequences of functional particles as:

Peter 6 *Will an a in there*

Any interpretation is really confusing.
There are cases where the presence of the **a** is clearly purely phonological. For example:

Peter 6 *more get* **a** *more*

Peter 6 a *ride a horsie*

Peter 5a *this thumb*

Peter 7 *hmmm my a*

[9]In the Veneziano & Sinclair's study, the chosen language is French, but we take as general some conclusions that seem to coincide with the observed phenomena in English acquisition

Beyond the non-definition of personal pronouns due to the weakness of the case system, we find the pronoun *I* as an *a*:

Peter 7 *a want milk*

Peter 7 *a want ta get out*

An interesting sequence of that reinforces our considerations is:

Peter 7 *a put it on (...) my put it on*

Finally, it is interesting to note the presence of elements that are, to some extent, a mixing between *a* and *to*: *ta*. The occurrence of this particle is rare and located explicitly at the very beginning of the funcional stage. The remarkable fact lies on the evidence that is located where it should be the preposition *to*. This could imply that in fact there is a transition from a pure phonological role to a functional one.

$a \rightarrow ta \rightarrow to$

Thus, we interpret *ta* as an intermediate stage but, due to its location within the sentence and the context, we assume it behaves as a preposition:

Peter 6 *[Have [ta [screw it]]$_{PP}$]*

Peter 7 *[Have[ta [screw it]]$_{PP}$]*

The emergence of English syntax is strongly tied to the emergence of functional particles. This is the reason why we decided to take into account this kind of lexical components: despite almost every utterance involving such items can be object of many considerations, there are enough motivation to try to define a descriptive criteria to deal with them.

4.2.5 Duplication of functional words

It is usual to find, at the beginning of the functional stage, that a verb that subcategorizes, for example, a prepositional phrase, display two successive prepositions:

Peter 6 *Look* **at in** *there*

To manage this kind of productions we assumed, first, that these imply that the child conceives[10] a syntactic structure that involves prepositional phrases.

[10] *Conceives* implies that the child is competent in this ind of productions, thus we are not using this verb in terms of explicit knowledge

Following this reasoning, we make the following structural description:

Look **at in** *there* → *[Look [at there]$_{PP}$]*

In that case, *in* is interpreted as an independent lexical item. Not only prepositions are involved in this duplication phenomena, but also determiners:

Peter 9 *One that screwdriver*

Interpretation rules out *one* as a member of any structure, leading the SD *[that screwdriver]$_{SD}$* alone.
A situation analogous to famous one described by Braine (p.160-161) [22] is:

Peter 7 *Get* **another** *one paper* → *Get another paper*

Thus, as above, determiner duplication is not considered in the structural analysis.

4.2.6 Non-structural lexical items

By this name, we designate the lexical items that are present in a conversational framework but cannot be explicitly interpreted as members of some syntactic structure, such as *Hello*, or *Ok*. The reason to include these elements as connected to the network is due mainly because their are produced in non-arbitrary context. Thus, we assume that they linked to the first element of the sentence they precede:

Peter 5 *Ok Patsy*

Obviously, previous reasons are at work when dealing with such items. Thus, the conversational context has to be analyzed to interpret these items. For example, In the following situation, *bye* has not been considered as a member of any structure. The reason is that it is produced among analogous expressions, leading it to interpret more in a pragmatic sense:

Peter 7 *see you, bye, see you*

Sequences of numbers or other elements produced as a list are not considered as members of any structure:

Peter 7 *one two three...*

Sometimes, personal nouns are produced by the child to demand attention from adult people. In these situations, we do not accept them as a members of structured sentences.

Some residual cases to be commented are the ones related with strings of nouns:

Peter 9 *Piece tape...*

Which are clearly unstructured, if conversational context does not conspire in the other way. Sequences like

Peter 9 *off on tv...*

are ruled out as structured ones because any structure proposal leads us to a certainly bizarre sentence in terms of meaning and because there is a lack of many elements that can act as clues to find some structure. Thus, they are considered as isolated lexical items:

Peter 7 An *Jenny*

4.2.7 Negation Structures

When the functional stage is being consolidated, we find more complex structures. Among others, interrogatives involving subject inversion or negation structures.
Negation structures sometimes imply the presence of the auxiliary *to do* are produced using the negative particle alone:

Peter 7 No *put it here*

This context suggests us that *No* could be replacing the auxiliary form *don't. Don't put it here* is its grammatical counterpart. Nevertheless, we consider *no* as replacing *don't* and, thus, as a member of a bigger syntactically structures utterance. Obviously, as we said above, context has to rule out interpretations such as **No***, put it here.* Analogous structures can be:

Peter 7 No *ride a bike*

There are other situations where we considered suitable not to consider *no* as a member of any structure:

Peter 8 *in the bag* **no**

We cannot conclude that there is a syntactic relation among the negation operator and some other lexical item of the string. Maybe a parametric [21] hypotheses could save this production by suggesting that the location of the negation operator within the structure may be a parametric feature. Despite interesting, we choose the rule these productions out for reliability purposes.

Furthermore, at the same time, there are productions like:

Peter 9 No *in this box*

suggesting that the child *knows* the ordering of negation structures in English. In the latter case, as before, we considered not risky to identify *no* as a member of a bigger structures sentence.

5 Corpora Annotation

Previous set of criteria enables us:

1. To identify and characterize syntactic structures in early child language.
2. To project them into word-word dependencies in order to annotate the corpus.

The corpora is annotated *by hand*. This enables us to be accurate and to manage ambiguous situations. The program used to perform the annotation is the so-called **Dependency Grammar Annotator** (DGA annotator). This program was developed by Marius Popescu [4] from the University of Bucaresti and has a nice and easy interface. It works with XML files, whose internal structure will be described in the example of the last section.

6 The average size of structures, $\langle \mathcal{S} \rangle$

With these criteria in hand, we are ready to perform a first analysis of grammar complexity. Such an analysis is closed to the classical MLU[11]. What we can compute, now is the average size of syntactic structures. Thus, in a production, we can, for example, find two syntactically unrelated structuress_1 and $s_2 s$. The number of lexical items of these structures will be its sizes $|s_1|$ and $|s_2|$. Such an utterance will contribute to the computation of $\langle \mathcal{S} \rangle$ with two structures.
For example, in

Look **at** **in** *that*

we have two structures:
$s_1 =$[Look, [at, that]] $\rightarrow$ $|s_1| = 3$
$s_2 =$in $\rightarrow$ $|s_2| = 1$
$\langle s \rangle = (3 + 1)/2 = 2$
(Note that single words are considered as size-1 structures) To obtain $\langle \mathcal{S} \rangle$ The average is computed over all utterances. Such a measure will provide us clues to decide whether the size of productions has information about grammatical complexity. Its evolution can be related with working memory limitations.

[11]Medium lenght of utterances, often measured on utterance size in words or morphemes

7 Building the Network

Once we analyzed a conversation, we can build the network. The process is as follows. Due to the nature of our analysis, we will have a collection of words, which define the set $\mathcal{W}$:

$$\mathcal{W} = \{car, it, ...\} = \{w_1, w_2, ..w_n\} \tag{1}$$

This define the set of nodes of our network. If, during the conversation, we find some structure where two words w_i, w_k are related syntactically -using the above criteria!- we say that $w_i rightarrow w_k$[12] and that there is a link $w_i \rightarrow w_k$.

$$\mathcal{E} = \{car \rightarrow want, it \rightarrow want, it \rightarrow fix...\} = \\ \{w_1 \rightarrow w_k, w_2 \rightarrow w_k, ...\} \tag{2}$$

Remark that:
All the words and links only appear once a time. This enables us to separate -as far as possible- some contextual deviations from the specific conversations. Also, there can be many isolated nodes.
Finally, we compute the adjacency matrix $\mathcal{A}_{ij}$. This matrix is the representation of the graph and the abstract object where all computations of graph complexity are performed. If the child produced n different words during the conversation, the size of this matrix will be, obviously n^2.
The adjacency matrix of the directed graph will be:

$$\mathcal{A}_{ij} = \begin{cases} 1 \leftrightarrow w_i \rightarrow w_i \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

If we consider the undirected version of this graph, $\mathcal{A}^u_{ij}$ will be defined as:

$$\mathcal{A}^u_{ij} = \begin{cases} 1 \leftrightarrow w_i \rightarrow w_j or w_j \rightarrow w_i \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

Note that $\mathcal{A}^u_{ij}$ is symmetrical, whereas $\mathcal{A}_{ij}$ it is not.
Now we are ready to perform an exhaustive analysis of network complexity.

7.1 Measures

A first and fundamental question we find when dealing with such measures is whether the net is made of a large number or small, isolated graphs or if it displays a clearly differentiated Giant Connected Component (GCC) that contains most of the connected words -i.e. words syntactically active in some production. The number of words contained on such a component or its relative size are interesting statistical indicators. Strikingly, from the very beginning, child's syntactic graphs display a clear and very differentiated GCC. For mathematical purposes, we will use the matrix representation of the connectivity pattern of

[12]Do not confuse it with the logical conditional

the GCC, the so-called *adjacency matrix*. An element of such a matrix is $a_{jk} = 1$ if there exists a link among the words W_j and W_k and $a_{jk} = 0$ otherwise. If the contrary is not indicated, -we will compute the following measures over the GCC of our graphs.
The number of links (or *degree*) $k_i = k(W_i)$ of a given word $W_i \in \mathcal{W}$ gives a measure of the number of (syntactic) relations existing between a word and its neighbors. The simplest global measure that can be defined on Ω is the average degree $\langle k \rangle$. For the T-th corpus, it will be defined as

$$\langle k \rangle_T = \frac{1}{N_w(T)} \sum_{W_i \in \mathcal{W}} k(W_i) \tag{5}$$

where $N_w(T)$ indicates the number of words present in the T-th corpus. This number is known to increase through acquisition in a steady manner. This and other measures are computed on the largest component of the graph.
Beyond the average degree, two basic measures can be used to characterize the graph structure of the GCC of the T-th corpus. These are the average path length (L_T) and the clustering coefficient (C_T). The first is defined as $L_T = \langle D_{min}(i, j) \rangle$ over all pairs $W_i, W_j \in \mathcal{W}$, where $D_{min}(i, j)$ indicates the length of the shortest path between two nodes. Roughly speaking, a short path length means that it is easy to reach a given word $W_i \in \mathcal{W}$ starting from another arbitrary word $W_j \in \mathcal{W}$. The second is defined as the probability that two vertices (e.g. words) that are neighbors of a given vertex are neighbors of each other. In order to compute the clustering, we define for each word W_i a neighborhood Γ_i. Each word $W_j \in \Gamma_i$ has been syntactically linked (via the above defined projection) at least once with W_i in some sentence. The words in Γ_i can also be linked among them, and it is what the clustering coefficient evaluates. The clustering $C(\Gamma_i)$ of this set is defined as

$$C(\Gamma_i) = \frac{1}{k_i(k_i - 1)} \sum_j \sum_{k \in \Gamma_i} a_{jk} \tag{6}$$

and the average clustering of the GCC concerning the T-th corpus is simply $C_T = \langle C(\Gamma_i) \rangle$. The clustering C provides a measure of the likelihood of having triangles in the graph. Concerning the average path length, for random graphs with Poissonian structure we have

$$D = 1 + \frac{\log [N/z_1]}{\log [z_2/z_1]} \tag{7}$$

being z_n the average number of neighbors at distance n. For Poissonian graphs, where $z_1 = \langle k \rangle$ and $z_2 = \langle k \rangle^2$, we have the following approximation: $D \approx \log n / \log \langle k \rangle >$. It is said that a network is a *small-world* when $D \approx D_{random}$ (and clearly $D \ll N$). The key difference between a Poissonian network and a real network is often $C \gg C_{random}$ [23].
Another quantity of interest is the degree of affinity among nodes with the same connectivity. In this way, the behavior of hubs is specially relevant, as well as

they organize the overall structure of the net. A network is said to be *assortative* if hubs tend to be connected among them. At the other side, a network is said to be *dissassortative* if hubs tend to avoid connections among them. Language networks at different scales display a high degree of dissassortativeness [13]. To quantify the degree of assortativeness, we use the so-called Pearson's coefficient for nets [24]:

$$\rho = \frac{c\sum_i j_i k_i - \left(c\sum_i \frac{1}{2}(j_i + k_i)\right)^2}{c\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left(c\sum_i \frac{1}{2}(j_i + k_i)\right)^2} \tag{8}$$

where j_i and k_i are the degrees of the edges at the ends of the ith edge with $i = 1, ..., m$, $c = \frac{1}{m}$ and being m the number of edges. If $\rho < 0$ the net is dissassortative, whereas if $\rho > 0$ the net is assortative.

8 Example

Below we have a fragment of the conversation transcribed in the Corpus *Peter 7*. We will detail the analysis that we perform. Firstly, we show the source corpus. We follow by selecting Peter's productions. After that we select the structures and analyze this structures and we tag them. We finish by computing $\langle \mathcal{S} \rangle$ of this fraction of text and by showing the obtained net.

8.1 The source

*PAT: hey Pete that's a nice new telephone looks like it must do
everything it must ring and talk and .
%mor: co—hey n:prop—Pete pro:dem—that v—be & 3S det—a adj—nice adj—new
n—telephone
n—look-PL v—like pro—it v:aux—must v—do pro:indef—everything pro—it
v:aux—must
v—ring conj:coo—and n—talk conj:coo—and .
%exp: Peter has a new toy telephone on table next to him
%com: ¡bef¿ untranscribed adult conversation
*CHI: xxx telephone go right there .
%mor: unk—xxx n—telephone v—go adv—right adv:loc—there .
%act: ¡bef¿ reaches out to lift phone receiver, pointing to place where
wire should connect receiver and telephone
*MOT: the wire .
%mor: det—the n—wire .
*PAT: oh ¡the & te¿ [//] the wire's gone ?
%mor: co—oh det—the n—wire v:aux—be & 3S v—go & PERF ?
%com: ¡aft¿ untranscribed adult conversation
*CHI: xxx need it my need it xxx .
%mor: unk—xxx v—need pro—it pro:poss:det—my n—need pro—it unk—xxx
.

%act: ¡aft¿ goes to his room on Mother's suggestion, returns with wire
*CHI: xxx .
%mor: unk—xxx .
*PAT: uhhuh .
%mor: co—uhhuh .
*LOI: why don't you bring your telephone down here Peter ?
%mor: adv:wh—why v:aux—do neg—not pro—you v—bring pro:poss:det—your n—telephone
adv—down adv:loc—here n:prop—Peter ?
*LOI: why don't you put it on the floor ?
%mor: adv:wh—why v:aux—do neg—not pro—you v—put & ZERO pro—it prep—on det—the n—floor ?
%act: ¡aft¿ Peter puts it on floor ¡aft¿ Peter is trying to attack "wire"
to phone and receiver
%com: ¡aft¿ untranscribed adult conversation
*LOI: what're you doing ?
%mor: pro:wh—what v—be & PRES pro—you part—do-PROG ?
*CHI: 0 .
%act: ¡aft¿ Peter goes to hall closet, tries to open it
*MOT: what do you need ?
%mor: pro:wh—what v—do pro—you v—need ?
*CHI: xxx .
%mor: unk—xxx .
(...)
*CHI: put in there .
%mor: v—put & ZERO prep—in adv:loc—there .
%act: attaching wire to phone
*LOI: ok it's all fixed oops it was out all fixed there .
%mor: co—ok pro—it v—be &3S qn—all part—fix-PERF co—oops pro—it v—be & PAST & 13S
adv—out qn—all v—fix-PAST adv:loc—there .

8.2 Selected Productions and Analysis

To work with the DGA Annotator, we need, firstly, to extract the child's productions. To do this, we programmed a routine in PERL language able to extract child's productions. Below there is a simple pseudocode as a sample:

```
FILE=PETER_k
for(i=5; i<=LONGFILE; i++)
{
   if(FILE[i]= /PETER/)
   {
   j=j+1;
   PETER[j]= "FILE[i]";
```

```
   }
}

for(i=0; i<=LONGFILE; i++)
{
      @PETER[i]= tr/*PETER:/ /;
      @PETER[i]=  tr/./ /;
      @PETER[i]=  tr/,/ /;
      @PETER[i]=  tr/;/ /;
      @PETER[i]=  tr/:/ /;
      @PETER[i]=  tr/!/ /;
      @PETER[i]=  tr/</ /;
      @PETER[i]=  tr/>/ /;
      @PETER[i]=  tr/?/ /;
      @PETER[i]=  tr/Â¿/ /;
      @PETER[i]=  tr/*/ /;
}
```

If we apply the above algorithm to the sample of text of the example, we obtain:

xxx telephone go right there
xxx need it my need it xxx
xxx
0
xxx
put in there

8.2.1 XML Format to be read by DGAanotator

Further we need to provide the obtained strings of words with a XML format, in order to manage them with the DGA Annotator. Below we have an example of the string *put in there*.

```
⟨?xml version="1.0" encoding="iso-8859-1"⟩
⟨!DOCTYPE DGAdoc SYSTEM "dga.dtd"⟩
⟨DGAdoc⟩
⟨s⟩
   ⟨tok⟩
       ⟨orth⟩put⟨/orth⟩
       ⟨ordno⟩1⟨/ordno⟩
   ⟨/tok⟩
   ⟨tok⟩
       ⟨orth⟩in⟨/orth⟩
       ⟨ordno⟩2⟨/ordno⟩
   ⟨/tok⟩
   ⟨tok⟩
```

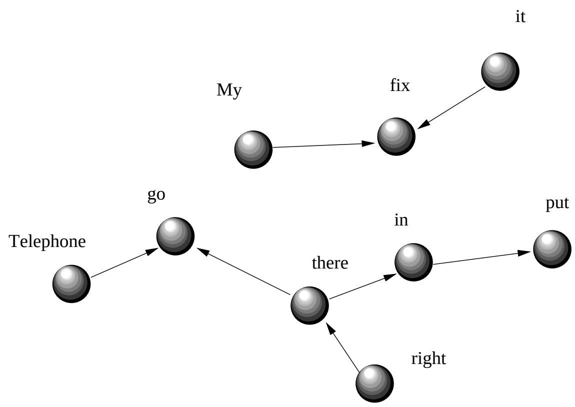

Figure 2: Graph of the sample.

```
        ⟨orth⟩there⟨/orth⟩
        ⟨ordno⟩3⟨/ordno⟩
    ⟨/tok⟩
⟨/s⟩
⟨/DGAdoc⟩
```

8.2.2 Selection of *valid* strings, annotation and computation of $\mathcal{S}$

We reject *xxx* and 0 as lexical items and proceed to annotate with the DGA anotator.

Once the corpus is annotated (with the criteria developed through the paper!) we generate the set of words. This can be done by sampling the XML file once annotated by using a routine close to the ones shown above (PERL or Python are the ideal languages). To compute graph parameters and more mathematical artifacts, it is a good choice to use a *stronger* language, such as C or C++.

$$\mathcal{W} = \{\text{telephone, go; right, there, need, it, my,} \\ \text{put, in}\} \tag{9}$$

And the analysis is, roughly speaking:

$$s_1 = [\text{telephone}[\text{go}[\text{right there}]_{\text{PP}}]_{\text{VP}}]_{\text{TP}} \;\; |s_1| = 4 \tag{10}$$

$$s_2 = [\text{need it}]_{\text{VP}} \;\; |s_2| = 2 \tag{11}$$

$$s_3 = [\text{my}[\text{need it}]_{\text{VP}}]_{\text{TP}} \;\; s_3 = 3 \tag{12}$$

$$s_4 = [\text{put}[\text{in there}]_{\text{PP}}]_{\text{VP}} \quad s_4 = 3 \tag{13}$$

Thus, we can compute $\langle \mathcal{S} \rangle$:

$$\langle \mathcal{S} \rangle = \frac{4+2+3+3}{4} = 3 \tag{14}$$

and, following the criteria developed above, we can define $\mathcal{E}$

$$\begin{aligned} \mathcal{E} = \{&\text{telephone} \to \text{go},\ \text{right} \to \text{here},\ \text{here} \to \text{go}; \\ &\text{it} \to \text{need},\ \text{my} \to \text{need},\ \text{there} \to \text{in},\ \text{in} \to \text{put}\} \end{aligned} \tag{15}$$

We have built the graph.

9 Acknowledgments

My indebtedness for Carlos Rodríguez-Caso for his ideas and his patient advising in teaching PERL. I also want to acknowledge Harold Fellermann for his help in programming python routines, Andreea Munteanu for the careful reading of the manuscript and Barbara Soriano for its useful comments on language acquisition process. My indebtness, also, to Ricard V. Solé and Sergi Valverde as well as this work is a part of a bigger work where their ideas have been illuminating. Finally, I'd like to acknowledge Ramon Ferrer i Cancho for his contributions at the beginning of the work.

References

[1] Melçuck (1989) *Dependency Syntax: Theory and Practice.* University of New York, New York

[2] Hudson, R. (2006). Language networks. The new word grammar. New York: Oxford University Press.

[3] MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

[4] Popescu, M. (2003) *Dependency Grammar Annotator*, Building Awareness in Language Technology (editors: Florentina Hristea & Marius Popescu). Editura Universităţii din Bucureşti, 2003, p.17-34.

[5] Chomsky, N. (1965) *Aspects of the theory of syntax.* MIT Press. Cambridge, MA.

[6] Chomsky, N. (1995) *The minimalist program.* MIT Press. Cambridge, MA.

[7] Chomsky N. (1963) Formal properties of grammars. In *Handbook of Mathematical Psychology* Vol II. Wiley Publishing Company, New York

[8] Chaitin, G. J. (1987) *Algorithmic information Theory*. Cambridge University Press. Cambridge, MA.

[9] Chomsky N. (1963) Formal properties of natural languages. In *Handbook of Mathematical Psychology* Vol II. Wiley Publishing Company, New York

[10] Zipf, G. K. (1949) *Human Behaviour and the Principle of Least-Effort* Addison-Wesley. Cambridge MA.

[11] Bollobàs, B. (1998) *Modern Graph Theory*. Springer. New York.

[12] Chomsky, N. & Miller, G. (1963) Finitary models of language users. In *Handbook of Mathematical Psychology* Vol II. Wiley Publishing Company, New York

[13] Ferrrer i Cancho, R., Solé, R. V. & Kohler, R. (2004) Patterns in syntactic dependency networks. Phys. Rev. E. 70 056135.

[14] Ferrer i Cancho, R. (2005) The structure of syntactic dependency networks: insights from recent advances in network theory. In *The problems of quantitative linguistics*, Altmann, G., Levickij, V. & Perebyinis, V. (eds.). Chernivtsi: Ruta. pp. 60-75.

[15] Radford, A. (1990) *Syntactic Theory and the Acquisition of English Syntax: the nature of early child grammars of English* Blackwell, Oxford.

[16] Bloom, L.; Hood, L. & Lightbown, P. (1974) Imitation in language development: If, when and why. Cognitive Psychology, 6, 380-420.

[17] Bloom, L.; Hood, L. & Hood, L. (1975) Structure and variation in child Language. Monographs of the society for Research in Child Development, 40, (serial No. 160)

[18] Sole, R. V., Corominas Murtra, B. & Valverde, S. *Unpublished Work*

[19] Veneziano, E. & Sinclair, H. (2000) The challenging status of filler syllabes on the way to grammatical morphemes.

[20] Bloom, L. (1970) *Language development: form and function in emerging grammars* (Research Monograph No. 59) The MIT Press, Journal of Child Language, 27, 461-500.

[21] Hyams, N. (1997) *Language Acquisition and the theory of Parameters.* Dordrecht: Reidel

[22] Braine, M. D. S. On two types of models of the internalization of grammars. In D. I.. Slobin (ed.), *The ontogenesis of grammar: a theoretical symposium.* Academic Press (New York)(1971) pp. 153-186

[23] Dorogovtsev, S. N. & Mendes, J. F. F. (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford University Press, New York.

[24] Newman, M. E. J.(2002) Assortative mixing in Networks. Phys. Rev. Lett. 89, 208701